STATISTICAL BENCHMARKING OF UTILITY SERVICE QUALITY

Mark Newton Lowry, Ph.D., *Partner*
Larry Kaufmann, Ph.D., *Vice President*
Donald Wyhowski, Ph.D., *Senior Economist*
Katherine Dresher, Ph.D., *Senior Economist*

November 9, 2000

PACIFIC ECONOMICS GROUP
22 E. Mifflin St., #302
Madison, WI 53703
Phone: 608.257.1522 Fax: 608.257.1540

INTRODUCTION

Service quality (SQ) benchmark mechanisms are common features of modern utility regulation. There is broad consensus that such mechanisms are needed to counterbalance the cost containment incentives generated by performance-based ratemaking (PBR) and the extended rate freezes that result from merger and restructuring agreements. SQ benchmark mechanisms are routinely used in Massachusetts rate freezes and PBR plans.

Despite their acknowledged importance, SQ benchmark mechanisms have not to date benefited from the same extensive theoretical and statistical research as have the cost-related PBR plan provisions. Absent such work, there remains a real risk that service quality provisions of PBR plans may not satisfy the just and reasonable standard for utility regulation.

Several regulatory Commissions have recently initiated generic proceedings on the design of SQ benchmark mechanisms. The Massachusetts Department of Telecommunications and Energy ("the Department") is currently conducting such a proceeding, which applies to jurisdictional energy distributors.[1] On August 17 of this year, it issued an Order proposing detailed provisions.

The Department proposes that distributors gather data on various SQ indicators and corresponding benchmarks. The proposed benchmark for each indicator for each company is an average of the recent historical values for the indicator for that company. The amount of annual data currently available for historical benchmark computations varies by indicator and company, and is in some cases as small as two or three.

The Department proposes to use comparisons between some particular performance indicators and their corresponding benchmarks as the basis for automatic revenue adjustments. Massachusetts law explicitly authorizes revenue penalties for poor quality performance under PBR plans in an amount up to 2% of a company's recent transmission and distribution revenue. The proposed penalty mechanisms would be

[1] Service Quality Standards, D.T.E. 99-84.

asymmetric in the sense that penalties could be levied for quality inferior to benchmarks whereas awards would not be levied for superior quality.

The penalty mechanisms would be non-linear in two respects. First, "deadbands" would be established in which quality levels nominally inferior to the benchmark would not be penalized. In the words of the Department,

> The proposed "deadband" recognizes the existence of normal statistical variations in service quality data, and provides a measure of protection to companies against being penalized for random statistical events.

The Department proposes that for each quality measure the deadband be established "equal to one standard deviation from the specific utility's historical average performance." The proposed penalty mechanism is also non-linear in that the relationship between the penalty and the performance comparison is parabolic in a range bounded by the deadband and the indicator value at which maximum penalties begin. The maximum penalty would be levied at a quality level two standard deviations from the mean historical performance.

A group of ten investor-owned local gas distribution companies and five investor-owned electric distribution companies has retained Pacific Economics Group to review the methodology proposed by the Department and to determine if that methodology is appropriate for meeting the Department's stated objectives. This paper is the report on our work to date. It considers how well-established tools of statistical science can be used to design SQ benchmark mechanisms. We devote particular attention to developing proper statistical methods for the calculation of deadbands.

Here is the plan for our report. Section 1 reviews the basic components of PBR benchmarking mechanisms, including those that apply to service quality. Section 2 introduces key concepts from statistical benchmarking and explains their use in SQ benchmark mechanism design. Precedents for statistical benchmarking in service quality regulation are discussed in Section 3. Our recommendations for service quality benchmarking in Massachusetts are presented in Section 4. A technical appendix provides the theoretical underpinnings for the discussion.

1. PBR BENCHMARKING MECHANISMS

1.1 Benchmarking Basics

A typical PBR benchmark mechanism consists of performance indicators, performance benchmarks, and an award/penalty mechanism. Performance indicators are measurable aspects of the company's operations that are monitored and evaluated. These indicators are variables in the sense that they can assume different values across companies and from period to period. Performance benchmarks are the numerical standards against which indicator values are compared.[2]

A penalty mechanism automatically adjusts a company's rates depending on the comparison of indicator values to the benchmark.[3] If the indicator values are inferior to the benchmark a penalty may be levied. Indicator values superior to the benchmarks may result in a reward. An auxiliary mechanism typically effects an adjustment to the company's rates to implement the penalty.

Penalty mechanisms can be applied in many ways. Penalty rates are commonly established for each indicator. A penalty rate determines the magnitude of the penalty per unit of deviation of the performance indicator and benchmark. The relationship between penalties and the deviation of indicators from their benchmark is frequently non-linear. For example, different penalty rates may apply for different deviations of the indicators from their benchmarks. Most notably, a deadband may exist in which deviations of indicator values from their benchmarks do not result in penalties.

Principles are needed for the development of PBR benchmark mechanisms. One is that a company should not be penalized for inferior performance when its performance is not inferior. We will refer to such error as a Type I error.[4]

---

[2] The indicators and benchmarks described above are sometimes referred to as metrics and standards.

[3] To streamline the discussion we will speak of award/penalty mechanisms as "penalty mechanisms" even though rewards may be realized as well as penalties. We may think of such rewards as "negative" penalties.

[4] There is a related concept called Type II error. That is the error that would result if the quality program were not judged to be inferior to the benchmark when it was.

Avoiding Type I errors is especially crucial when benchmark mechanisms are of the asymmetric, penalty-only variety. With symmetric penalty mechanisms, Type I errors still occur but tend to balance out over time between incorrect determinations of superior and inferior performance. The evaluation mechanism is thus, on balance, fair. With asymmetric mechanisms there are no such offsets and the chance of the mechanism penalizing companies unfairly over the years is greater.

How might a company be falsely judged to have inferior SQ performance? To answer this question, it is important to understand the process by which values for performance indicators are generated. Performance indicators depend in part on actions undertaken by a company and in part on external factors. External factors are those factors that affect a company's operations but are not controlled by company personnel.

Such factors differ across companies, and change over time for each individual company. Some are volatile in the sense that they are prone to fluctuations that are hard to predict. If benchmark mechanisms do not take account of differences in external factors between companies and over time the chance of a Type I error increases.

PBR benchmarking can take account of external factors in several ways. One is to choose performance indicators in such a way that fluctuations in business conditions are less likely to lead to Type 1 errors. An example is to exclude observations for periods in which there are highly unusual external factors.

Another general approach is to choose benchmarks that reflect the impact of external factors on indicators. This is most commonly done by basing benchmarks on a company's own historical values of the performance indicators. This approach ensures that benchmarks will reflect the *typical* external factors faced by the company, which may differ substantially between companies. However, it will not control for local fluctuations in external factors around their norms.

External factors can also be accommodated through the design of penalty mechanisms. Suppose, by way of example, that the value of a performance indicator for a company is known to fluctuate over time in a certain range due to changes in external factors. If the benchmark is the mean value of this indicator for a recent historical period it will then be difficult to establish with certainty that indicator values in this range reflect inferior performance on the part of the company even if the indicator value is nominally

inferior to the benchmark. Deadbands can reduce the likelihood of a Type I error due to fluctuations of indicators within their normal range.

1.2 Application to Distribution Service Quality

These benchmarking basics have a ready application in the design of SQ provisions of PBR plans. Service quality is the single most common dimension of operations subject to benchmark PBR mechanisms. SQ benchmark mechanisms typically consist of quality indicators, quality benchmarks, and a penalty mechanism.

The risk of a Type I error is substantial in quality benchmarking. The quality of energy distribution service is potentially influenced by a number of external factors, which may be called quality "drivers". The list of relevant factors includes weather (*e.g.* winds, lightning, extreme heat and cold), vegetation (contact with power lines), the amount of undergrounding mandated by local authorities, the degree of ruralization in the territory (typically increasing the exposure of feeders to the elements and lengthening response times when faults occur), the difficulty of the terrain served, the mix of residential, commercial, and industrial customers, the incidence of poverty, the heterogeneity of languages spoken, the rate of growth in the number of customers, the tendency of customers to relocate, and regulatory changes such as a restructuring of the industry to promote competition. These factors vary substantially between distributors and some are quite volatile.

The level of service quality does not depend solely on quality drivers. Rather, it depends on the efficacy of a distributor's service quality effort. We may refer to this informally as quality effort. A quality effort encompasses quality-related work practices, worker training, and capital investments. Relevant work practices include the size of call center staffing and power line maintenance procedures such as tree trimming. Relevant capital investments include the size and sophistication of call center communications equipment and software.

In developing a quality effort, it is rational from both a shareholder and customer perspective to balance considerations of cost and quality. It is generally not cost effective to have the same quality levels in service territories with markedly different quality drivers. For example, few will argue that rural power distributors should have the same SAIFI numbers as an urban distributor because of the higher cost required. It is,

similarly, not in general cost effective for a distributor to hold its quality level constant in the face of local fluctuations in drivers. As one example, it is not cost effective to ensure that power distribution quality remains high even during severe storms. As another, it is not cost effective for call centers to be staffed so as to maintain call response quality even when there is an unusual surge in calls. As the saying goes, you don't size a church that you are building to house the Easter Sunday crowd.

The rational balancing of cost and external factor considerations has the result that the level of quality provided by distributors varies across companies. Furthermore, distribution systems are rationally designed to deliver fluctuating quality levels. If not adjusted, indicator values can be especially sensitive to the incidence of extreme conditions that systems are not designed to accommodate.

Our discussion suggests that fluctuations in the values of quality indicators are often due to fluctuations in quality drivers rather than change in quality effort. Differences in indicator values between companies are, analogously, often due to differences in the drivers they face and not differences in their effort. Both phenomena raise the risk of a Type I error in SQ benchmarking.

Suppose, by way of example, that the quality benchmark is the mean of recent historical values of the quality indicator. Due to fluctuations in external factors, the quality of service during a PBR year may then be inferior to that in recent years despite no change in the company's quality effort. Quality benchmark mechanisms should be designed to reduce Type I error by isolating the change in a company's quality effort.

All of the methods described above for taking account of external factors can be used in SQ benchmarking. Some of these tools are recognized, at least in principle, by commissions and reflected in the design of SQ benchmark mechanisms. To begin, quality indicators in approved plans are often measured in ways that exclude the impact of unusual quality drivers. A prime example is that SAIFI and SAIDI indicators typically exclude outages that occur during periods of major storms.[5]

SQ benchmarks are, secondly, routinely designed to reflect what is known about quality drivers and their effect on quality. This is most commonly done by basing

[5] However, the definition of what constitutes a major storm varies from state to state and can even vary among utilities within a state.

benchmarks on company-specific historical indicator values, as the DTE proposes.[6] A distributor's historical quality norms such as mean values of the indicators for recent years presumably reflect local quality drivers, which may differ greatly from those facing other utilities.

Note, however, that SQ benchmarks based on historical averages reflect the quality drivers only during the years of the historical period. The drivers operative during the individual PBR plan years may differ greatly from these norms. Were utilities to ensure that they were not improperly penalized for quality inferior to recent averages, they would then have to upgrade their quality effort to a level *superior* to that which generated the benchmark. This would involve considerable cost which might ultimately be paid by customers.

External business conditions affecting quality can also be handled through the design of penalty mechanisms. This is most commonly achieved in SQ benchmark plans by the use of deadbands. Deadbands reduce the likelihood that a reduction in quality due to more adverse quality drivers is confused with a deterioration of quality effort. In placing SQ benchmarking mechanisms on a scientific foundation, a key question is the proper deadband width to achieve this goal.

[6] Company-specific historic standards are also useful to the extent that the goal of service quality provisions is to prevent a deterioration of service quality.

2. STATISTICAL BENCHMARKING

2.1 General Principles

Well-established statistical methods are available for use in the design of SQ benchmark mechanisms. These mechanisms provide two major benefits. One is the ability to systematically account for the inherent variation in quality levels due to quality drivers and thereby better identify change in quality effort and reduce the chance of Type I errors. A second benefit is the ability to control the likelihood of a Type I error and thereby set at a level acceptable to policymakers. The use of statistics to construct benchmarks for economic performance indicators is a branch of economic science called statistical benchmarking.

The use of statistics to design SQ benchmark mechanisms begins by viewing quality indicators as random variables. A random variable is a variable whose values are drawn from a population characterized by a probability distribution. The value assumed by the indicator in each period is then viewed as a draw from the population. A collection of historical values of an indicator is then a random sample. By way of example, SAIDI may be viewed as a random variable that may in a given year and for a given quality effort assume a range of values depending on the state of quality drivers in that year.

The distribution of a random variable can be represented mathematically. The mathematical formula may contain such parameters as the mean and the standard deviation of the distribution. The mean of a quality indicator, which is the expected level of quality, is a measure of its central tendency. The actual value of the indicator will usually differ from its mean. The standard deviation of an indicator's distribution measures its "spread", that is, the tendency of indicator values to differ from their mean.

We may reasonably posit that the mean of the indicator is a stochastic function of a set of measurable quality drivers and the level of quality effort. The expected level of

quality then differs for each possible set of drivers.[7] It will decline if there is a decline in quality effort.

If we do not measure the relationship between quality indicators and quality drivers, we may still posit that a quality indicator is a random variable with a mean that is an unknown function of quality drivers and the level of quality effort. Actual values of the indicator may differ from the mean when quality drivers differ from their norms.

We can calculate the mean and standard deviation of a sample of historic values of a quality indicator. These statistics, properly called the sample mean and sample standard deviation, are also random variables in as much as they are calculated from random variables. Their values are specific to the external factors in play during the sample period. We may view the sample mean and sample standard deviation as estimates of the true mean and standard deviation of the population. Statistical theory permits us to measure the accuracy of these estimates. For example, the estimates will typically be more accurate the larger is the size of the sample.

Suppose, now, that we take the difference between the value of a quality indicator and the sample mean of the indicator for a recent pre-plan historical period. Our analysis suggests that this difference accurately compares the level of quality effort before and after the new plan begins if the net effect of external factors during the plan year is the same as the net effect of external factors during the sample period. To the extent that the net effects of external factors differ before and during the plan, the likelihood of a Type I error increases. Intuition suggests that the chance for a Type I error is especially great if the Company is penalized for a quality level that is inferior to the benchmark but well within the usual range of fluctuation for that indicator.

Statistical theory can be used to assess whether indicator values during the PBR period are drawn from the same population as the indicator values used to construct the benchmark. Granted some assumptions regarding the distribution of the performance indicator, we can construct hypothesis tests regarding values of a quality indicator during the years of a PBR plan. For example, we can test the hypothesis that its value in a given

[7] The parameters of this function could in principle be estimated using historical data. The function could then be used to predict a value for the performance indicator given local values for the business condition variables in the model. The research required to develop reliable statistical models of this kind is, however, very much in its infancy.

plan year is drawn from the same population as the benchmark. If it is not, a change in quality effort may be indicated.

The hypothesis test can take the form of comparing a test statistic to a certain critical value. The critical value is determined from what is known about the distribution of the test statistic and the degree of confidence placed on the hypothesis test. The confidence level on which the test statistic is based is the likelihood of a Type I error. The confidence levels most widely used in statistical research on economic variables are 95% and 99%.

An equivalent and more intuitive approach to hypothesis testing is to examine whether the value of the indicator during a plan year is bounded by a confidence interval constructed from the sample mean and the test statistic. If it is, we cannot reject the hypothesis that the value is drawn from the same population as the historic sample. If the quality level is inferior to the most inferior value in the confidence interval we can conclude that the difference between the new value and the benchmark is statistically significant. In this event, the company's quality may be deemed to be a significantly inferior quality performer.

As confidence intervals widen, a determination of significantly superior performance is less likely. The formula for the test statistic used to construct the confidence interval can be examined to determine which factors affect the width of the interval. Confidence intervals will be wider the greater is the confidence level assigned to the hypothesis test, the larger is the sample standard deviation, and the smaller is the size of the sample from which the estimate of the mean is constructed.[8]

It is interesting to note how the determinants of a proper confidence interval differ from the fixed standard deviation approach that the Department proposes for deadband calculation. The confidence interval does include the sample standard deviation because this is highly germane for determining how much quality typically fluctuates around the mean due, presumably, to fluctuating quality drivers. The deadband is constructed using a confidence level since this controls the chance of Type I error. Confidence interval

[8] In the event that we have posited the mean as a function of measurable business conditions, statistical theory suggests that the confidence interval will also be wider to the extent that included business condition variables fail to explain the variation in sampled values and the business conditions faced by the company deviate from the mean values of these variables in the sample.

construction also requires the size of the sample to control for the degree of uncertainty surrounding our estimates of the mean and standard deviation of the benchmark population.

An important step in the construction of a hypothesis test for a performance indicator is the selection of the test statistic. To the extent that the performance indicator is normally distributed the "t" test statistic has desirable statistical properties. Notable among these is its suitability for small sample sizes. The t-statistic is one of the most widely used test statistics in economic research.

Statistical theory can be used to estimate the frequency of Type I error resulting from a fixed standard deviation approach to deadband construction such as that proposed by the Department. This will vary with the size of the sample. Assuming that the quality indicator is normally distributed, for example the probability of error if the deadband is set at one sample standard deviation from the sample mean is 18.3% when ten years of historical data are available. The likelihood of Type I error using the Department's proposed approach is higher the smaller is the sample size used to calculate the mean. For a sample of only two historic values, for instance, the likelihood of Type I error is 28.2%.

3. PRECEDENTS FOR STATISTICAL BENCHMARKING IN REGULATION

There is ample precedent for the use of statistical benchmarking in SQ regulation. Most notable are the statistical tests approved for use in telecommunications. Under the Telecommunications Act of 1996, regional Bell Operating Companies (BOCs) can enter in-region, interLATA toll markets if they can demonstrate that they have opened their traditionally non-competitive local exchange and exchange access markets to competing carriers. A key issue in evaluating how far BOCs have gone in opening their markets is the quality of "wholesale" services provided by the BOCs to competing local exchange carriers (CLECs).

Verizon, an incumbent local exchange carrier and regional Bell Operating Company (BOC), filed for authorization to provide interLATA long distance service in Massachusetts and New York State.[9] It proposed similar comprehensive performance enforcement mechanisms in these states that would be activated were the company authorized by the Federal Communications Commission (FCC) to provide in region inter LATA services. The mechanisms were approved by the New York Public Service Commission (NYPSC) and the Massachusetts DTE.

The approved mechanisms include Performance Assurance Plans ("PAPs") that feature an automatic process under which CLECs receive bill credits in the event that Verizon fails to satisfy predetermined quality standards established for a large number of indicators.[10] A typical benchmark mechanism compares the quality of service offered by Verizon to CLECs to the quality contained in Verizon's retail services. The Verizon enforcement mechanism for New York and Massachusetts relies heavily on formal hypothesis tests that involve test statistics. The methodology was developed through a collaborative process that included the affected customers (CLECs) as well as Verizon.

The test statistics used to construct the confidence intervals vary with the nature of the performance indicator. A t statistic is used for many indicators for which the sample size is less than thirty. The chosen confidence level is 95%.

[9] The Verizon operating unit in New York State was then called Bell Atlantic-New York.

[10] See NYPSC Case 97-C-0139 (November 3, 1999) and Massachusetts DTE 99-271 (September 5, 2000).

In its order approving the PAP, the NYPSC acknowledges the role that hypothesis testing can play in deadband calculation. It states that, "the PAP was designed…to produce minimum…scores that provide a 95% level of confidence that BA-NY will not be unfairly held accountable for substandard scores that result from random variation."[11]

The NYPSC also acknowledges that an asymmetric penalty-only mechanism increases the importance of controls for Type I error, stating that

> The objective of the PAP's statistical framework was to not hold BA-NY responsible for random variation given that the company is not rewarded for its good service.[12]

The Department has also endorsed the general approach to deadband determination contained in the Verizon plan. In its order approving the PAP for Massachusetts states that

> The Department finds that adopting and maintaining a 95% confidence interval protects carriers from any likelihood of financial consequences for peer performance… A 95 percent confidence level is generally accepted as an adequate statistical standard.[13]

The FCC, long a PBR innovator, has since authorized Verizon to commence interLATA service in New York.[14] [15] This was the first such authorization it granted. In its order, the FCC cites the enforcement mechanism as key evidence that Verizon has opened its system to competition.[16]

The FCC decision for New York includes an extensive discussion of the statistical methodology in the plan. It acknowledges that "random factors outside the control of the

---

[11] See NYPC's *op cit* p16.

[12] *ibid* p17.

[13] DTE 99-271 (September 5, 2000) p26.

[14] Verizon's petition for Massachusetts is pending with the FCC. Application by Verizon New England, Inc., Bell Atlantic Communications, Inc. (d/b/a Verizon Long Distance), NYNEX Long Distance Company (d/b/a Verizon Enterprise Solutions), and Verizon Global Networks, Inc., for Authorization Under Section 271 of the Telecommunications Act of 1996 to Provide In-Region, InterLATA Services in Massachusetts, CC Docket No. 00-176.

[15] *Memorandum Opinion and Order in the Matter of Application by Bell Atlantic New York for Authorization Under Section 271 of the Communications Act to Provide In-Region InterLATA Service in the State of New York*, CC Docket No, 99-295, December 22, 1999, p. 212.

[16] "We find that the performance monitoring and enforcement mechanisms in place in New York, in combination with other factors, provide strong assurance that the local market will remain open after Bell Atlantic receives section 271 approval."

BOC" influence the values of performance indicators and could cause an indicator to assume a value inferior to the benchmark even though its quality effort is equally effective. Because of such factors, the FCC states that "the use of statistical analysis to take account of random variation in the metrics is desirable."[17] The FCC explains that statistical analysis involves the treatment of performance indicators as random variables. In the Commission's words,

> Statisticians would say that the [performance indicator] observations are a sample taken from the population. The population is the theoretical set of values obtained if an infinite number of observations were taken of the underlying process. Therefore the population mean is the theoretical mean produced by the process, while the sample mean is the measured mean.[18]

The appropriate application of statistical theory to the construction of deadbands would involve hypothesis tests based on test statistics. The FCC explicitly approved the use of a t statistic to construct confidence intervals for small sample sizes.[19] It also approved the use of a 95% confidence level for the determination of statistical significance. In other words, the FCC assumed that the quality of a service dimension covered by an indicator is satisfactory unless the indicator value is significantly different from the benchmark value using a hypothesis test based on a t-stat (or other appropriate test statistic) and a 95% confidence interval. The FCC states that,

> We use the 95% confidence interval because it is a commonly used standard, and because it gives us a reasonable chance of detecting variations in performance not due to random chance, with few false conclusions that variations are not due to random chance.

This general approach to carrier-to-carrier service quality regulation is now spreading to states outside the northeast. The FCC has now authorized Southwestern Bell Telephone (SWBT) to provide interLATA service in Texas.[20] The SWBT application includes a Performance Remedy Plan that was previously approved by the Public Utility

[17] FCC *op cit* Appendix B.

[18] *ibid* Appendix B.

[19] According to the FCC decision, no commenter in the FCC proceeding opposed the use of the Verizon test statistics.

[20] Memorandum Opinion and Order in the Matter of Application by SBC Communications etal Pursuant to Section 271 of the Telecommunications Act of 1996 to Provide In-Region, InterLATA Services in Texas, CC Docket No. 00-65 (June 30, 2000).

Commission of Texas. That plan features a number of mechanisms for benchmarking the quality of service provided by Southwestern Bell to CLECs and for exacting penalties as appropriate.

As in the Verizon plan, the mechanisms involve hypothesis tests based on test statistics and thus are equivalent to using deadbands based on confidence intervals constructed from the test statistics. A 95% confidence level is employed. Once again, the test statistics evolved from a collaborative process in which SWBT and the CLECs both had input. The test statistics employed depend on the character of the indicator and the size of the sample available.

Other states that have not yet adopted rigorous statistical benchmarking procedures for service quality have considered a fixed standard deviation approach for the establishment of service quality standards and have opted for a broader deadband than that proposed by the DTE. A case in point is Pennsylvania where the Public Utility Commission (PAPUC) has issued a rulemaking on reliability of electric service. It has fixed deadbands at two standard deviations from the sample mean, reasoning that

> This methodology will produce reasonable and realistically achievable initial minimum performance standards which reflect the variability of historic reliability performance.[21]

[21] Final Order, Docket No. M-00991220, 29 April 1999.

4. SERVICE QUALITY BENCHMARKING RECOMMENDATIONS

In light of this discussion, we recommend the following general guidelines for the development of more scientific SQ benchmark mechanisms in Massachusetts.

1. Each penalty mechanism should incorporate a deadband based on a confidence interval. We propose that confidence intervals be calculated using t-statistics. As we have seen, these are useful in samples of small size. Deadbands based on a confidence interval constructed from the t-statistic will reflect the standard deviation of the data used in benchmark construction, as in the Department's proposal. Unlike the Department's mechanism, it will also adjust automatically for the size of the sample and the desired confidence level. A deadband so designed will lead to penalties only if a distributor's quality is significantly inferior at the chosen confidence level. A sensible deadband would allow for a 5% probability of Type I error, as in the Performance Assurance Plans approved by this Department, the NYPSC, and the FCC for Verizon.

2. Care must be taken to develop samples of adequate size for hypothesis testing. One means of achieving this is to recompute the sample mean benchmarks routinely as new data are accumulated until samples of adequate size are realized. Consideration should also be paid to postponing the activation of penalty mechanisms when samples of minimally acceptable size are unavailable. For example, there exists strong theoretical arguments suggesting a minimum of three and possibly four sample points are needed to establish a valid deadband.

3. Performance indicators are more likely to assume a normal distribution when the underlying data are normalized to exclude identifiable business conditions, such as severe weather, which are known to lead to asymmetrically distributed quality outliers. Exclusion of outliers should also be considered for customer service measures. For example, we might throw out call center observations for periods

of severe storms or unusually large call demand.

4. With regard to the penalty mechanism, we propose to modify the Department's proposed parabolic mechanism such that the standard deviation is replaced by the width of the confidence interval as the basis for deadband construction. The proposed formula[22] is

$$Penalty = \left[0.25 \left(\frac{Observed\ Result - Historical\ Average\ Result}{Confidence\ Interval} \right)^2 \right] \cdot Maximum\ Penalty$$

The Technical Appendix presents an application of the recommended procedures to the case of a single quality indicator: the lost time due to accidents at Boston Gas. The company has 10 years of data on this indicator. The mean value of the indicator is 1.587. The sample standard deviation is 0.656, while two standard deviations is 1.312. Using the Department's single standard deviation approach, the upper band on the deadband is 2.243. The maximum penalty is incurred at 2.899. Using the alternative statistical benchmarking approach, and a t-statistic, the upper bound on the deadband is 2.848. The value where the maximum penalty is incurred is 3.528.

5. To further reduce the risk of Type I error the penalty mechanism can be modified by basing penalties on each company's net quality performance over a multi-year period. Under the current proposal, the companies are never rewarded but are subject to penalties in any year that performance on an indicator falls below the lower band. The plan can be modified to allow for the "banking" of significantly superior performance in any given year to offset findings of significantly inferior performance in other years. The company is then penalized only if its performance is significantly inferior on balance. For example, each company can be judged ultimately on its performance over the full five year period of the PBR plan rather than year by year. Even if no awards are offered for performance that is on balance superior, this will improve the fairness of the plan by creating the chance to offset penalties attributable to Type I error.

TECHNICAL APPENDIX

In service quality benchmark mechanisms, a penalty due to year-to-year fluctuations in external factors is an example of what statisticians' term "Type I" error. The Type I error can be illustrated by considering a person on trial for murder. In this case, a Type I error would occur if a person were convicted of a murder they did not commit. Although decreasing Type I error would cause an increase in Type II error, namely the probability of allowing a guilty person to be acquitted, philosophies of jurisprudence as well as conventions in hypothesis testing from the theory of statistics suggest that maintaining a very low Type I error is preferable.[23]

The purpose of this appendix is to provide an assessment of the Type I error associated with various penalty mechanisms. We begin by discussing the general framework for evaluating Type I errors for a penalty mechanism when a benchmark based on the utility's historical performance is used. Next, we assess the level of Type I error associated with the one standard deviation approach for establishing the "deadband" as proposed in D.T.E. order 99-84. We then propose an alternative "test statistic" approach to determining a deadband based on the concept of a confidence interval for an out-of-sample forecast. Finally, we illustrate by way of an example how a deadband is computed using the test statistic approach.

Framework for the Analysis

We start by letting Q denote a particular quality indicator. We suppose that the realized as well as the yet to be realized values of Q represent independent draws from an identical population having a normal distribution with mean μ and variance σ^2. We denote this distribution by $N(\mu, \sigma^2)$.

[22] Penalty does not apply until the utility exceeds the bounds of the deadband.

[23] We will not argue here as to the "correct" level of type I error that is "fair" but only say that in statistics probabilities with magnitudes of 10%, 5% or 1% are most common. There is also ample precedence in opinions and decisions made by State Regulatory Commissions suggesting a 5% level type I error for determining deadbands.

For ease of exposition we will model this random variable by letting $Q = \mu + \varepsilon$ where ε denotes the (unobservable) error term capturing the year-to-year fluctuations in external factors that are beyond the company's control. This random error is assumed to have a standard normal distribution, which we denote by $N(0, \sigma^2)$. Since external factor are deemed unobservable variables in the present setting, the error term can be viewed as being driven by the net effect of all such variables from their normal values. The presence of those unobservable drivers may lead to systematic variation in ε and hence violate the normality assumption for ε. We set aside this potential complication until later in the discussion.

The focus of this discussion is the difference between the historical average benchmark for a and its future realization, Q. This average historical benchmark, or sample mean, for measured metrics is defined to be the sum of the observations divided by the number of observations, or sample size, T.[24] Therefore, the level of Type I error of interest is simply the probability of the out-of-sample forecast error given by $(Q - \hat{\mu})$ taking on values beyond some proposed value for its deadband. To compute this probability will require some technical details contained in the following lemma.[25]

Lemma Assuming the random sample given by $(Q_1, \ldots, Q_T)$ and the yet to be realized random variable Q are draws from the population $N(\mu, \sigma^2)$[26], the following statements are true.

(i) $\hat{\mu} \sim N\left(\mu, \sigma^2 / T\right)$ where $\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} Q_t$

[24] Mathematically, if we denote the sample mean by $\hat{\mu}$ then $\hat{\mu} = \frac{1}{T}\sum Q_t$.

[25] The authors upon request will provide proofs of these results.

[26] The assumption of normality is maintained throughout this appendix. Given the small size of the samples involved, it is not feasible to test this assumption. Should the companies sometime in the future, when a larger sample size is available, find that their data suggest the assumption of normality inappropriate, alternatives to the procedures discussed here can be considered.

(ii) $E[\hat{\sigma}_2] = \sigma^2$ where $\hat{\sigma}^2 = \sum_{t=1}^{T} \frac{(Q_i - \hat{\mu})^2}{T-1}$

(iii) $E[Q - \hat{\mu}] = 0$

(iv) $\text{var}(Q - \hat{\mu}) = \sigma^2 + \frac{\sigma^2}{T}$

Statements (i) and (ii) are the standard properties for the sample mean and variance. Statement (iii) says that the sample mean $\hat{\mu}$ provides an unbiased forecast for Q. The statement (iv) contains the variance of the forecast error and illustrates the two sources of uncertainty inherent in any out-of-sample forecast. The first term σ^2 represents the uncertainty in not knowing the future realization of Q. The second term σ^2/T represents the variation of the estimated benchmark from its true value μ.

The Standard Deviation Approach

Under the guidelines proposed in D.T.E. 99-84 a utility would incur a penalty if a particular SQ measure deviates more than one standard deviation above its historical average. We will refer to this scheme for establishing a deadband as the "fixed standard deviation" (SD) approach. The Type I error associated with this rule is given in the following proposition.

Proposition 1: Assuming the random sample given by (Q_1, …, Q_T) and the yet to be realized random variable Q are draws from the population N(μ, σ^2)[27], it follows that $P(Q > \hat{\mu} + \hat{\sigma}) = P\left(t > \sqrt{\frac{T}{T+1}}\right)$ where t is a random variable having a student-t distribution with (T-1) degrees of freedom.

[27] See footnote 24.

Two remarks about the level of Type I error imposed on the utilities are in order. The first concerns the magnitudes established under the proposed guidelines. For example, for a utility with ten years of historical data (*i.e.* the sample size proposed in the plan), the probability that Q is more than one standard deviation above its sample mean is the area under the curve of a student-t distribution (with nine degrees of freedom) to the right of the value 0.95. Using statistical software[28] that contains cumulative distribution functions we compute the level of Type I error to be 18.3%. The second more striking remark is that the level of Type I error increases as sample size (T) decreases. This is a result of the increased uncertainty of whether or not the estimated sample mean is close to its actual counterpart (*i.e.* μ). For example, the level of Type I error when the number of available historical data is five years rises to 20.7%. For a utility with the minimum two years of data the level of Type I error is 28.2%.

Test-Statistic Approach

An alternative and more scientific approach to establishing a deadband can be developed based on the concept of hypothesis testing that is firmly ground in the theories of probability and statistics. This approach assumes that an acceptable frequency of Type I error (say α) has already been established. It then determines the appropriate value for the deadband taking into account the number of years of historical data a utility has available for a particular SQ measure. The formula for this deadband is provided by the following proposition.

Proposition 2: Assuming the random sample given by (Q_1, …, Q_T) and the yet to be realized random variable Q are draws from the population $N(\mu, \sigma^2)$[29], it follows that $P\left(Q > \hat{\mu} + t_\alpha \hat{\sigma} \sqrt{1+\frac{1}{T}} \right) = \alpha$ where t_α is the value of the abscissa obtained from a student-t distribution with (T-1) degrees of freedom.

[28] The Pacific Economics Group uses the mathematical and statistical software package Gauss-386i, a product of Aptech Systems, Inc.

[29] See footnote 24.

Two comments are again in order. First, this approach allows the level of Type I error to be controlled regardless of the amount of historical data (*i.e.* sample size T) a utility has available. This is accomplished by increasing the deadband for indicators with fewer observations. Thus, for example, an indicator for which there are two years of historical data[30] will have a deadband more than three and a half times that of an indicator for which there are ten years of data, everything else held constant. The rationale is straightforward. Since there is less reliability in the sample mean as an estimator of the true population mean, a larger deadband must be set to compensate for this increased uncertainty.

Second, the deadband given in proposition 2 includes the deadband given in the fixed SD approach as a special case. To see this we must choose a 15.9% level of Type I error so that t_α takes on a value one. Then we must assume that the sample of historical data becomes that of the underlying population in the sense of letting the sample size T increase without limit. In this very unrealistic scenario, the deadband in proposition 2 reduces to the deadband in the SD approach.

An Example Computing Deadbands and Upper Band

We use the following data for the Safety Performance measure Lost work time due to accidents to illustrate by way of an example how a deadband is computed using the test-statistic approach. We also compute the upper band at which point a penalty would be imposed on a utility if a particular service quality (SQ) indicator exceeds the value of this upper band in some future year. For purpose of comparison we will also compute the values for these same measures using the standard deviation approach.

Table 1

Lost Work-Time Accident Rate

Year	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991
Q	1.32	1.07	1.08	1.50	0.70	1.10	2.60	2.30	1.80	2.40

[30] It should be noted that this test statistic may be inappropriate for sample sizes smaller than four since the first two months do not exist for a sample size of two and the second moment does not exist for a sample size of three.

Source: Boston Gas Historical Data

Using the data found in Table 1 we compute the following values:

Sample size (T): 10
Sample mean $(\hat{\mu})$: 1.587
Sample standard deviation $(\hat{\sigma})$: 0.656

Using the following Table we find the value for a level of Type I error of 5% (i.e. $\alpha = .05$) and sample size ten (i.e. T = 10) of the abscissas from a student-t distribution to equal 1.833.

Table 2

Abscissas from t-Distribution for a given level of Type I error α and sample size T.

T[31]	$t_{.05}$	$t_{.01}$
2	6.314	31.821
3	2.920	6.965
4	2.353	4.541
5	2.132	3.747
6	2.015	3.365
7	1.943	3.143
8	1.895	2.998
9	1.860	2.896
10	1.833	2.821

Source: This table is based on Table 12 of *Biometrika Tables for Staticitians*. Volume I, edited by E. S. Pearson and H. O. Harley (1970).

Using the above information we begin by computing the value of the upper band using the standard deviation approach and find it to be $(\hat{\mu} + \hat{\sigma})$ = 1.587 + 0.656 = 2.243.

We next compute value of the width of the deadband (DB) using the test-statistic approach: DB = $t_{.05} \cdot \hat{\sigma} \cdot \sqrt{1+\frac{1}{T}}$ = (1.833)(0.656)(1.049) = 1.261. The upper band using this approach is then found to be ($\hat{\mu}$ + DB) = 1.587 + 1.261 = 2.848.

[31] The given value for sample size T corresponds to (T-1) degrees of freedom.

Using the value for the deadband under the test-statistic approach computed above, we note that none of the observed values of the measures in the sample exceed the historical mean (*i.e.* 1.587) by more than the deadband. That is, none of the observed values in the sample exceed the value of the upper band 2.848. On the other hand, under the standard deviation approach, the observed measures exceed the computed value of the upper band of 2.243 three out of the ten years.

A second (hypothetical) example illustrates the effect of using a smaller sample on the size of the deadband, everything else held constant. To this end, suppose T = 2 while the sample mean and standard deviation continue to take the values $\hat{\mu}$ = 1.587 and $\hat{\sigma}$ = 0.656, respectively. From Table 2 we see that $t_{.05}$ now takes the value 6.314. Computing the deadband for this smaller sample size we find it has now increased to DB = $t_{.05} \cdot \hat{\sigma} \cdot \sqrt{1+\frac{1}{T}}$ = (6.314)(0.656)(1.225) = 5.074 or more than four times its magnitude when the sample size was ten.